ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 27 за 2013 г.

Борисов Л.А., Орлов Ю.Н., Осминин К.П.

Идентификация автора текста по распределению частот буквосочетаний

Рекомендуемая форма библиографической ссылки: Борисов Л.А., Орлов Ю.Н., Осминин К.П. Идентификация автора текста по распределению частот буквосочетаний // Препринты ИПМ им. М.В.Келдыша. 2013. № 27. 26 с. URL: http://library.keldysh.ru/preprint.asp?id=2013-27

Ордена Ленина
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Российской академии наук

Л.А. Борисов, Ю.Н. Орлов, К.П. Осминин

Идентификация автора текста по распределению частот буквосочетаний

Москва — 2013

Борисов Л.А., Орлов Ю.Н., Осминин К.П.

Идентификация автора текста по распределению частот буквосочетаний

Исследованы распределения расстояний между распределениями триграмм, получена оценка точности частот буквосочетаний в зависимости от длины текста и даны оценки вероятности правильной идентификации автора текста по близости текста к его средневзвешенному эталону в смысле распределения частот. Построены авторские длины представительности для большого числа писателей и показано, что стабилизация триграмм происходит примерно на половине текста независимо от автора и длины текста. Проведен анализ литературного наследия Е.И. Рерих с целью кластеризации ее произведений и проверки ряда утверждений о возможном соавторстве.

Ключевые слова: эмпирическая вероятность, минимально достаточная длина текста, идентификация автора

Borisov L.A., Orlov Yu.N., Osminin K.P.

Identification of a text author by the letter frequency empirical distribution

The distances distributions between empirical triplet distributions are investigated. The accuracy estimation of these distributions is obtained depending on the length of the text. The method of author identification is examined on the broad class of literature texts. The stabilization length of triplet distributions is approximately equal to one half of the text without dependence on author and text length. The example of cluster method is given for E.I. Roerich philosophical texts.

Key words: empirical probability, minimal text length, author identification

Работа выполнена при поддержке гранта РФФИ, проект № 13-01-00617

Содержание

1. Введение

Вопросы статистического определения автора литературного текста постоянно находятся в сфере интересов как литературоведов, так и математиков. Практическую важность имеют аспекты исторические, правовые, культурологические. Вызывает интерес и само исследование творческой работы мозга с точки зрения алгоритмичности этого процесса. Также имеют самостоятельную ценность и статистические методы анализа, развитые применительно к таким многомерным по своим атрибутам объектам, как литературные тексты, написанные профессиональными писателями. Последнее подразумевает, что писатель в своих произведениях придерживается определенной манеры письма, что и позволяет использовать статистические методы для определения принадлежащих ему текстов.

Настоящая работа продолжает исследования, промежуточный итог которых был подведен в монографии [1]. В [1] для определения автора текста использовался метод близости частот встречаемости букв и пар букв (диграмм). Здесь мы опишем результаты анализа частот триграмм и обсудим некоторые аспекты точности идентификации атрибутов текста. Подчеркнем, что анализируется только последовательность букв того алфавита, на котором написан текст. Пробелы, знаки препинания, цифры, буквы другого алфавита и иные символы игнорируются.

Задача-максимум состоит в том, чтобы, механически разобрав текст на определенные элементы (в данном случае буквосочетания), указать его местоположение в многомерном «фазовом» пространстве мировой литературы: язык, эпоха написания, литературное направление (течение, принадлежность к определенному кружку и т.п.), тип (проза или поэзия), формат (роман, повесть, очерк, эссе), жанр (детектив, фантастика, триллер и т.д.), и, наконец, автор. Это означает, что в идеале должны быть определены проекторы на каждое подпространство в той текстовой структуре, которая сопоставляется отдельному произведению и позволяет с достаточной точностью провести анализ. Такой текстовой структурой в нашем методе является плотность функции распределения текста по буквосочетаниям или *n*-ПФР, где *n* отвечает порядку *n*-грамм. Хотя для дискретных распределений термин «ПФР» как плотность распределения не вполне удачный, он показался авторам более приемлемым, чем, например, «эмпирическая частотная характеристика» текста (ЭЧХ).

Мы не беремся утверждать, что поставленная выше задача-максимум может быть полностью решена только методом математической статистики. Наша цель – расширить возможности данного метода с чисто описательных в данной области знания и придать ему статус исследовательского инструмента, который он имеет в естественных науках, и который может быть использован в практически важных задачах литературоведения. Из всего многообразия вопросов мы сосредоточимся далее на двух: идентификация автора

неизвестного текста и кластеризация произведений, близких по своим 3-ПФР в определенной норме.

Первая задача формулируется следующим образом. Есть некая библиотека текстов без авторов и прочих классификационных атрибутов. Требуется наилучшим образом, т.е. с наименьшей ошибкой, распределить произведения по авторам и жанрам, не определяя сами эти жанры и авторов. Иными словами, требуется дать ответ примерно в такой форме: в данной библиотеке определенные n_{11} текстов написаны одним автором в одном жанре, n_{12} текстов написаны им же, но в другом жанре, и т.д., n_{21} текстов написаны вторым автором в первом жанре, n_{22} текстов во втором жанре и т.д. Это – пример задачи многомерной кластеризации.

Вторая задача – это собственно идентификация автора и (или) жанра. Она решается путем сравнения текстов с набором «эталонов», составляющих библиотеку. Все тексты библиотеки известны как элементы фазового пространства. Требуется внутри известной структуры определить наиболее вероятное место для одного текста с неизвестными атрибутами. Как вариант – можно снабдить ответ экспертной оценкой того, что подходящего эталона в библиотеке нет. В данной работе мы подробно рассмотрим вторую задачу и приведем некоторые примеры решения первой.

Предполагая различать авторов с помощью некоторого «механического» принципа, который основан на сопоставлении тексту определенной математической конструкции – например, функции распределения, – и нахождении расстояния между двумя такими конструкциями в подходящей норме, мы должны иметь для обоснования этой идеи достаточное количество конкретных примеров. Подчеркнем, что строго математического доказательства такого принципа быть не может, поскольку автор имеет право и возможность написать текст так, что тот будет гарантированно отличаться от всех его предыдущих произведений. Для этого достаточно наложить на творческий процесс ограничения формы: пусть, например, все слова в тексте начинаются с одной и той же определенной буквы, или каждое последующее слово начинается с той буквы, которой окончилось предыдущее, и т.п. Мы будем рассматривать только такие произведения, в которых нет нарочитых условий на правила соединения слов в предложения, и для которых форма вторична по отношению к содержанию.

Основное предположение состоит в том, что текст, написанный автором с целью передачи смысловой информации, содержит в себе «проекцию» индивидуальности мышления автора. Тексты, написанные одним и тем же автором, должны обладать близкими «проекциями», а разными авторами – заметно различаться, так что, если удастся подобрать подходящее правило сравнения «проекций» и, отфильтровав то общее, что присуще всем вообще «людям пишущим», можно формально отличить одного автора от другого.

Наилучшая различающая мера в пространстве «проекций» априори не известна, поскольку следует надеяться, что писательский труд – процесс творческий, а не алгоритмический. Возможно, что мера, близкая к оптимальной, обнаружится в ходе перебора достаточно большого количества разных вариантов сравнения расстояний между текстами. В [1] были рассмотрены разные нормы, в которых можно сравнить между собой 1-ПФР и 2-ПФР. Наилучшая точность идентификации автора была получена в норме суммируемых функций: ошибка идентификации по 1-ПФР составила 0,15, а по 2-ПФР 0,05 для совокупности «30 авторов – 300 текстов». В настоящей работе мы пошли несколько дальше в техническом плане и изучили близость между распределением триграмм, т.е. между 3-ПФР в той же норме. Авторы произведений в аналогичном эксперименте были определены безошибочно. Идентификация авторов по 4-ПФР привела к ошибке 0,07. Следовательно, возникла задача объяснения «феномена 3-ПФР» как наиболее точного инструмента. В настоящей работе подробно обсуждается доверительный уровень получаемых статистических выводов, а также вопрос о том, какой объем текста достаточен для достижения требуемой точности в оценке близости между n-ПФР.

2. Распределение расстояний между выборочными распределениями

Формализуем задачу идентификации автора неизвестного текста. Она состоит в следующем. Имеется библиотека, содержащая тексты, представленные в виде ПФР (однобуквенных или многобуквенных) для A известных авторов. Пусть K_a – имеющееся количество текстов a-го автора, и $N_{i,a}$ – количество букв в i-ом тексте этого автора, $i = 1,2,...,K_a$. Длина каждого из текстов достаточна для проведения статистического анализа (см. далее п. 3.). Обозначим $f_{i,a}(j)$ n-ПФР соответствующего текста, где аргумент j меняется от 1 до $\alpha(n) = 33^n$. Для каждого автора определим его средневзвешенную ПФР:

$$F_a(j) = \frac{1}{N_a}\sum_{i=1}^{K_a} f_{i,a}(j) N_{i,a}, \quad N_a = \sum_{i=1}^{K_a} N_{i,a}. \qquad (1)$$

Эти ПФР будут играть в дальнейшем роль авторских эталонов.

В (1) мы для краткости записи формул пренебрегли единицей по сравнению с N_a при подсчете пар букв в тексте для 2-ПФР (или двойкой при подсчете триграмм, если речь идет о 3-ПФР и т.д.), т.к. $N_a >> n$.

Введем «библиотечную норму» ρ_{ik} как расстояние между ПФР текстов i и k в норме суммируемых функций:

$$\rho_{ik} = \| f_i - f_k \| = \sum_{j=1}^{\alpha(n)} | f_i(j) - f_k(j) |. \qquad (2)$$

Для каждого автора a построим плотность функции распределения $g_a^+(\rho)$ отклонений $\rho_{i_a,a}$ «его» текстов, а также распределение $g_a^-(\rho)$ отклонений $\rho_{k_b,a}$ «чужих» произведений от его средней ПФР F_a. Обозначим $G_a^{\pm}(\rho)$ соответствующие интегральные функции распределения. Минимальное значение ρ, при котором $G_a^+(\rho)=1$, обозначим ρ_a^+, а максимальное значение ρ, при котором $G_a^-(\rho)=0$, обозначим ρ_a^-. Смысл введенных величин в том, что все ПФР текстов автора a находятся на расстоянии не более ρ_a^+ от его средней ПФР (1), и аналогично все ПФР других авторов находятся от нее на расстоянии не менее ρ_a^-. Величина $1-G_a^+\left(\rho_a^-\right)$ есть вероятность ошибочно признать за произведение автора «a» чужой текст (ошибка второго рода), а величина $G_a^-\left(\rho_a^+\right)$ есть вероятность ошибочно отвергнуть произведение автора «a», посчитав его за чужое (ошибка первого рода). Назовем расстоянием разделения авторов такое значение $\hat{\rho}$, для которого ошибка идентификации автора текста минимальна:

$$\hat{\rho}=\arg\min\left(1-G^+(\rho)+G^-(\rho)\right)=\arg\max\left(G^+(\rho)-G^-(\rho)\right). \qquad (3)$$

Эта величина может служить верхним уровнем для кластеризации текстов.

На рис. 1 приведены плотности $g^{\pm}(\rho)$ распределений расстояний для 3-ПФР по достаточно большой совокупности авторов (30 авторов, 300 текстов) без деления авторов на разные «лица», т.е. без кластеризации их произведений по принципу взаимной близости.

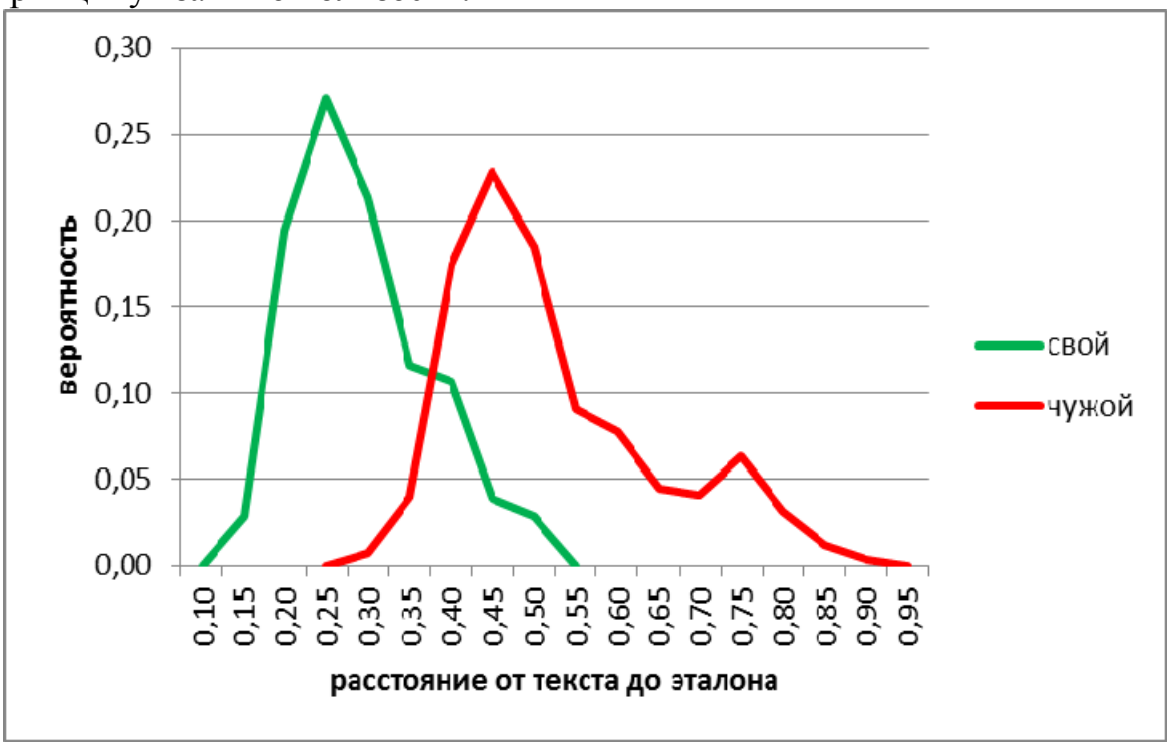

Рис. 1. Распределение расстояний между 3-ПФР текста и авторского эталона

Естественно, на момент сравнения «свое» произведение исключалось из эталона (1), так что ПФР i-го произведения a-го автора сравнивалась с его квази-эталоном:

$$F'_{i,a}(j) = \frac{1}{1 - N_{i,a}/N_a}\left(F_a(j) - f_{i,a}(j)\frac{N_{i,a}}{N_a}\right), \qquad (4)$$

$$\left\| f_{i,a} - F'_{i,a} \right\| = \frac{1}{1 - N_{i,a}/N_a} \sum_{j=1}^{\alpha(n)} \left| f_{i,a}(j) - F_a(j) \right| = \frac{\left\| f_{i,a} - F_a \right\|}{1 - N_{i,a}/N_a}. \qquad (5)$$

Те авторы, которые пишут примерно в одном стиле (например, Д. Донцова, Б. Акунин), имеют узкое распределение $g_a^+(\rho)$, а авторы, пишущие в разных жанрах (такие как Л. Толстой, Н. Гоголь), имеют более широкое распределение расстояний от своих текстов до авторского эталона. Распознать последних авторов можно только в том случае, если расстояния от их текстов до чужих эталонов заметно превосходят характерное расстояние до своего эталона. Нам важно сравнить эти распределения в целом по группе разных авторов. Распределение расстояний между авторскими эталонами показано на рис. 2. Аналогичный вид имеют распределения для эталонов 2-ПФР и 1-ПФР.

Рис. 2. Распределение расстояний между 3-ПФР авторских эталонов

Это распределение имеет выраженный правый скос и по типу ближе к гамма-распределению. Заметим, что абсцисса максимума распределения на рис. 2 примерно совпадает с абсциссой точки пересечения графиков «свой-чужой» на рис. 1. На рис. 3 показано распределение расстояний между отдельными текстами своих и чужих авторов.

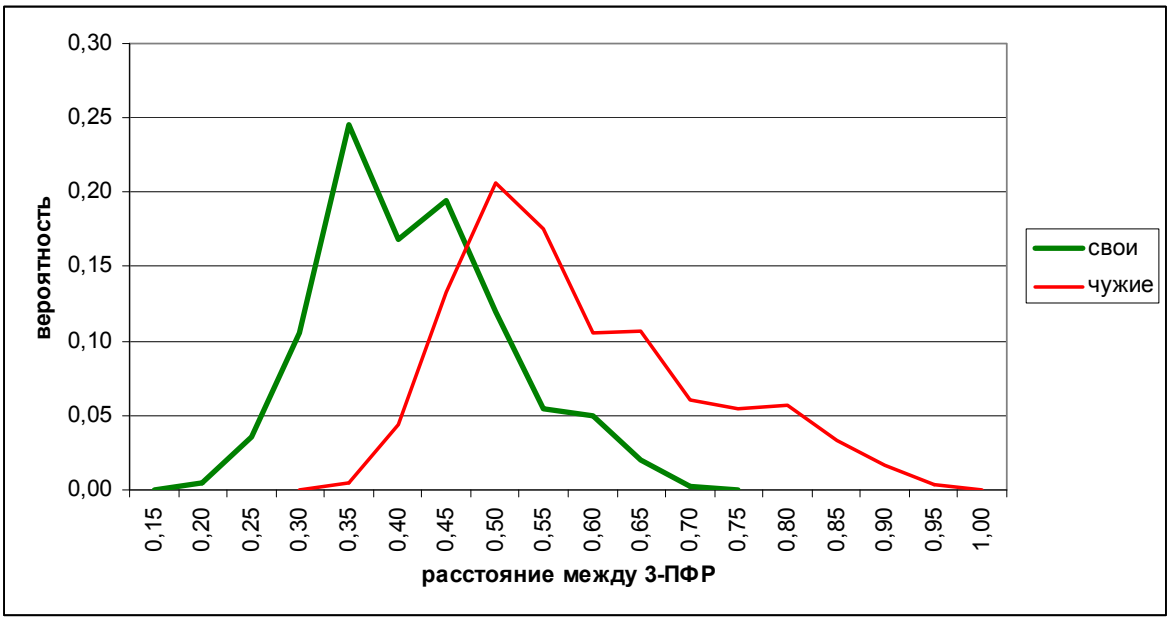

Рис. 3. Распределение расстояний между 3-ПФР отдельных текстов

Из рис. 1-3 видно, что характерное расстояние между авторскими эталонами примерно такое же, как и между текстами одного автора, и, в свою очередь, в 2 раза больше, чем расстояние от текста до своего эталона. Последнее обстоятельство позволяет трактовать эталоны как различные генеральные совокупности. Кроме того, из рис. 3 видно, что между отдельными текстами одного и того же автора расстояние может быть весьма большим (правая часть зеленого графика), тогда как зеленый график на рис. 1 более узкий. Следовательно, эталонное сравнение более точно позволит определить автора, чем сравнение между собой отдельных текстов.

Некоторые характеристики распределений рис.1 приведены в табл. 1.

Табл. 1. Характеристики распределений расстояний между текстом и эталоном

Показатель	1-ПФР	2-ПФР	3-ПФР
Среднее значение l_s (свой текст)	0,04	0,13	0,27
Среднее значение l_d (чужой текст)	0,08	0,23	0,50
Стандартное отклонение σ_s (свой текст)	0,02	0,04	0,08
Стандартное отклонение σ_d (чужой текст)	0,03	0,07	0,12
Расстояние разделения $\hat{\rho}$	0,05	0,19	0,35
$\rho+$	0,11	0,30	0,55
$\rho-$	0,03	0,10	0,25
Вероятность ошибки I: $G^-(\rho+)$	0,47	0,65	0,72
Вероятность ошибки II: $1-G^+(\rho-)$	0,72	0,57	0,50
Общая площадь под распределениями	0,38	0,35	0,22
Интервал 0,90 нормировки (свой текст)	0,03-0,06	0,08-0,16	0,20-0,40

Пусть теперь имеется текст «0» неизвестного автора, который надо идентифицировать внутри данной библиотеки. Автором текста «0» считается тот из авторов «*a*», для которого норма $\rho_a^0 = \| f_0 - F_a \|$ разности между ПФР $f_0(j)$ текста «0» и средней авторской ПФР $F_a(j)$ минимальна:

$$\rho_a^0 = \| f_0 - F_a \|, \quad a^0 = \arg\min_a \rho_a^0 . \tag{6}$$

Правило (6) применяется только в том случае, если $\min_a \rho_a^0 \le \hat{\rho}$. Если же оказалось, что $\min_a \rho_a^0 > \hat{\rho}$, т.е. минимальное расстояние превосходит длину разделения, то принимается решение об отсутствии в библиотеке подходящего авторского эталона.

Описанный метод оказался весьма точным в определении автора неизвестного текста и эффективным при программной реализации. Чтобы иметь представление о его априорной погрешности, следует оценить минимальный объем текста, который гарантирует, что статистическая неопределенность в эмпирической оценке ПФР как идеализированной генеральной совокупности будет не выше заданного уровня.

3. Точность оценки вероятностей буквосочетаний

Если рассматривать отдельное произведение или его фрагмент как выборку из некоторой «генеральной совокупности», которую своим творчеством реализует конкретный писатель, то следует определить, с какой точностью оцениваются генеральные вероятности по эмпирической *n*-ПФР.

Пусть $F^{(n)}(j)$ есть генеральная совокупность, за которую примем авторский эталон (1), а $f^{(n)}(j;N)$ есть ПФР отдельного произведения или отрывка. Верхний индекс n отвечает размерности ПФР.

Если бы процесс появления букв в тексте был полностью случаен (что, очевидно, не так), то статистика

$$t = \sqrt{N-1}\,\frac{\left| f^{(n)}(j;N) - F^{(n)}(j) \right|}{s(j;N)} \tag{7}$$

имела бы распределение Стьюдента [2] с $N-1$ степенью свободы. Здесь $s^2(j;N)$ есть выборочная дисперсия частоты, равная

$$s^2(j;N) = f^{(n)}(j;N) \cdot \left(1 - f^{(n)}(j;N)\right). \tag{8}$$

Поскольку $N >> 1$, то вместо квантиля распределения Стьюдента в оценке доверительного интервала для вероятности $f^{(n)}(j;N)$ можно взять квантиль нормального распределения $u_{1-\varepsilon/2}$, где ε – уровень значимости, на котором

принимается решение о близости распределений. Тогда в приближении $N-1\approx N$ на уровне значимости ε выражение $\left|f^{(n)}(j;N)-F^{(n)}(j)\right|$ не превосходит величины

$$s(j;N)u_{1-\varepsilon/2}/\sqrt{N}\,.$$

Следующее требование является ключевым. Поскольку уровень значимости не должен быть точнее интегрального уровня неопределенности в позиционировании доверительного интервала $\left|f^{(n)}(j;N)-F^{(n)}(j)\right|$, то естественно потребовать выполнения условия

$$\sum_{j=1}^{\alpha(n)}\left|f^{(n)}(j;N)-F^{(n)}(j)\right|\le\varepsilon\,. \tag{9}$$

Поэтому, если будет выполнено условие

$$\left|f^{(n)}(j;N)-F^{(n)}(j)\right|\le\varepsilon F^{(n)}(j)\,, \tag{10}$$

то при условии $u_{1-\varepsilon/2}\dfrac{s(j;N)}{\sqrt{N}}\le\varepsilon F^{(n)}(j)$ будет достигнут требуемый уровень значимости для статистики (7). Однако если некоторые вероятности в результате выбранного разбиения сами оказались малы, много меньше ε, то нет необходимости требовать, чтобы и они были оценены с той же точностью. Поэтому уместно для каждой вероятности выбрать свою точность аппроксимации ε_j и считать, что требуемый в целом уровень значимости определяется средневзвешенной по разбиению точностью, так что

$$u_{1-\varepsilon/2}=\frac{1}{\Sigma_N(n)}\sum_{j=1}^{\alpha(n)}s(j;N)u_{1-\varepsilon_j/2}\,, \tag{11}$$

где сумма, определяющая влияние мелкости разбиения гистограммы на точность оценки эмпирических вероятностей, равна

$$\Sigma_N(n)=\sum_{j=1}^{\alpha(n)}s(j;N)=\sum_{i=1}^{\alpha(n)}\sqrt{f^{(n)}(j;N)\cdot\left(1-f^{(n)}(j;N)\right)}. \tag{12}$$

Тогда из (10, 11) получаем, что

$$\sum_{j=1}^{\alpha(n)}u_{1-\varepsilon_j/2}\frac{s(j;N)}{\sqrt{N}}=u_{1-\varepsilon/2}\frac{\Sigma_N(n)}{\sqrt{N}}\le\varepsilon\sum_{j=1}^{\alpha(n)}F^{(n)}(j)=\varepsilon\,, \tag{13}$$

откуда на уровне значимости ε следует оценка

$$\frac{u_{1-\varepsilon/2}}{\varepsilon}\le\frac{\sqrt{N}}{\Sigma_N(n)}\,. \tag{14}$$

При заданной точности ε и числе $\alpha(n)$ состояний формула (14) для знака равенства дает оценку на минимальную длину выборки, при которой эта

точность достигается в среднем. Поскольку функция $u_{1-\varepsilon}$ табулирована (см., напр., [2]), то функция $u_{1-\varepsilon}/\varepsilon$ известна. Некоторые ее значения приведены в [3]. При $\varepsilon > 0{,}01$ имеет место следующая аппроксимация [4] квантиля нормального распределения:

$$u_{1-\varepsilon/2} = \sqrt{-\frac{\pi}{2}\ln\left(1-(1-\varepsilon)^2\right)}. \qquad (15)$$

Относительная ошибка аппроксимации (15) составляет 0,037.

Функция $u_{1-\varepsilon}/\varepsilon$ монотонно убывает с ростом ε, поэтому к ней существует обратная, значение которой и дает верхнюю оценку точности определения эмпирических вероятностей в зависимости от количества состояний $\alpha(n)$. Обозначим для краткости

$$\varphi(\varepsilon) = \frac{u_{1-\varepsilon}}{\varepsilon}, \quad \psi = \varphi^{-1}, \quad z \equiv z(N,n) = \frac{\sqrt{N}}{\Sigma_N(n)}. \qquad (16)$$

Тогда точность оценки n-ПФР определяется формулой

$$\varepsilon = 2\psi(2z). \qquad (17)$$

Сумма (12) весьма слабо зависит от длины выборки, но существенно меняется при изменении размерности ПФР. Округленные средние значения $\Sigma_N(n)$ по библиотеке в 300 текстов приведены в табл. 2.

Табл. 2. Значения $\Sigma_N(n)$ и точность оценки n-ПФР

	1-ПФР	2-ПФР	3-ПФР	4-ПФР
$\Sigma_N(n)$	5	18	51	105
Значения ε по формуле (17)				
N = 10 тыс.	0,08	0,22	0,40	0,60
N = 30 тыс.	0,05	0,15	0,30	0,42
N = 50 тыс.	0,04	0,12	0,25	0,39
N = 100 тыс.	0,03	0,10	0,20	0,31
N = 500 тыс.	0,02	0,05	0,15	0,20
N = 1 млн	0,01	0,04	0,10	0,15

Обращаясь к табл. 1, сравним точность определения n-ПФР и разность между средними расстояниями до «своего» и «чужого» эталона. Для того чтобы корректно распознавать своего и чужого авторов, требуется точность в оценке вероятностей ПФР не хуже указанного расстояния разделения. Для 1-ПФР эта величина равна 0,04, а из табл. 2 находим, что минимальная длина текста равна 50 тыс. знаков. Для 2-ПФР разность между средними расстояниями равна 0,10, а для 3-ПФР 0,23, что требует длину текста на уровне 100 тыс. знаков.

Тем не менее, как показывают конкретные примеры распознавания автора текста (см. далее п. 5), даже тексты длиной 10 тыс. знаков идентифицируются

методом 3-ПФР безошибочно, чего не могло бы быть, если бы тексты разных авторов просто представляли собой случайную последовательность букв из соответствующих генеральных совокупностей. Следовательно, возможность распознавания автора текста лежит не только в статистических свойствах распределений текста по буквам. То, что тексты своего автора неизменно оказываются ближе к своему эталону, чем к чужому, связано, видимо, с тем, что автор воспроизводит свою генеральную совокупность точнее, чем это «предписывается» просто случайным процессом.

Для того чтобы выяснить, насколько обоснованно это предположение, надо определить, как быстро стабилизируется ПФР фрагмента произведения к ПФР всего текста.

4. Авторская длина представительности

Введем понятие длины ε-стационарности текста как такой минимальной длины $L(\varepsilon)$, что для любого фрагмента этого текста большей длины $L' > L(\varepsilon)$ выполняется условие

$$\sum_{j=1}^{\alpha(n)} \left| f(j;L') - F(j) \right| \le \varepsilon . \qquad (18)$$

Здесь $F(j)$ есть n-ПФР всего произведения. Пусть длина всего текста составляет N знаков. Тогда легко показать, что при длине фрагмента текста $k \ge N_\varepsilon$, где

$$N_\varepsilon = \left[(1-\varepsilon) N \right] \qquad (19)$$

и квадратные скобки обозначают целую часть числа, условие (18) заведомо выполнено. Таким образом, при длинах выборок, мало отличающихся от полной длины текста, наблюдается линейная поточечная (т.е. для каждой буквы) сходимость выборочного распределения к распределению всего текста. Если бы авторы имели длину стационарности своих произведений вблизи этой верхней оценки $L(\varepsilon) \approx N_\varepsilon$, то корректно сравнивать произведения разных длин было бы невозможно, а тогда и задача распознавания автора текста по функции распределения символов потеряла бы смысл. Интерес представляют ситуации, когда $L(\varepsilon)$ существенно меньше, чем N_ε при таких значениях ε, которые отвечают точности, достаточной для различения авторов. В таком случае с достоверностью $1-\varepsilon$ можно считать, что на длине $L(\varepsilon)$ распределение становится неотличимо от распределения всей выборки.

Формализуем критерий сравнения между собой ПФР текстов, имеющих заметно различающиеся длины. Рассмотрим, например, 1-ПФР двух текстов, длины которых равны N_1 и N_2, а расстояние между ними в норме (2) равно $\rho_{12} = \left\| f_{N_1} - f_{N_2} \right\|$. Это расстояние корректно отражает различие между текстами только в том случае, если больший объем имеет длину стабилизации

$L(\lambda)$ на некотором уровне λ, существенно меньшем, чем само ρ_{12}, причем $L(\lambda)$ не превосходит длины меньшего из текстов.

Термин «существенно меньше» означает следующее. Пусть имеется K текстов одного автора, и L_0 есть длина минимального из них. Для k-го текста этой длине отвечает определенный уровень квазистационарности ε_k:

$$\varepsilon_k = \max_j \sum_{i=1}^{n} \left| f_{L_0}^{(k)}(i;j) - f_{\max}^{(k)}(i) \right|. \qquad (20)$$

Если положить $\lambda = \max_k \varepsilon_k$, то каждый текст на длине L_0 будет λ-стационарным. Рассмотрим $K(K-1)/2$ попарных расстояний ρ_{ij} между текстами одного автора. Пусть распределение этих расстояний имеет среднее $\bar{\rho}$ и дисперсию σ^2. Зададим точность δ, с которой мы предполагаем различать тексты. Именно, если $1-\delta$-квантиль эмпирического распределения попарных расстояний больше, чем расстояние между данным неизвестным текстом и любым из по крайней мере $[(1-\delta)K]$ базовых текстов, то этот текст с доверительной вероятностью $1-\delta$ будем считать принадлежащим перу того же автора. Этот вывод корректен, если только $\lambda < \delta$. Кроме того, если оказалось, что $\lambda > \sigma/\bar{\rho}$, то сам автор пишет настолько разнообразно, что его нельзя точно идентифицировать. Такие авторы могут представлять собой контрпримеры к статистическому методу идентификации. Следовательно, точность метода определяется долей плохо идентифицируемых авторов в выборке текстов, и потому характеризует не только метод, но и саму выборку.

Таким образом, первоочередной задачей, связанной с длиной стационарности $L(\varepsilon)$ текста, является установление того, насколько велик разброс этой величины в зависимости от ε по совокупности произведений одного автора или, как вариант, по совокупности произведений близких длин. Будем называть далее функцию $L(\varepsilon) = L_a(\varepsilon)$ длиной представительности автора «*a*», показывающей, на какой характерной длине текста набирается распределение, близкое к ПФР всего произведения в смысле (18).

На рис. 4-6 показаны примеры длин представительности как функций уровня ε-стационарности текстов для произведений Ю. Никитина, В. Пелевина и В. Шишкова. Выбранные тексты значительно различаются по объему, и, чтобы иметь возможность корректно сравнивать их ПФР, требуется уверенность в том, что достаточная статистика набирается уже на минимальном из текстов. Для сравнения приведена также кривая длины, отвечающей заданной точности оценки 3-ПФР в соответствии с формулой (17). Она показана на графике как сплошная черная линия с легендой «stat set».

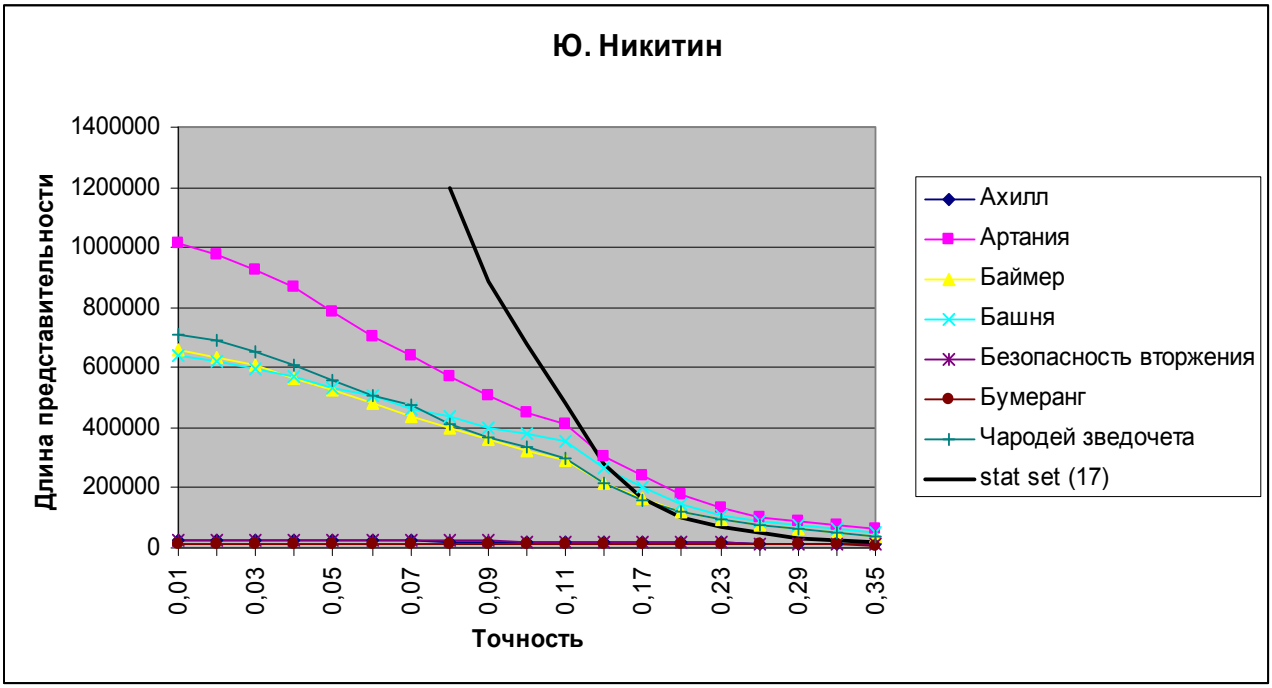

Рис. 4. Зависимость длины представительности от уровня стационарности для произведений Ю. Никитина, 3-ПФР

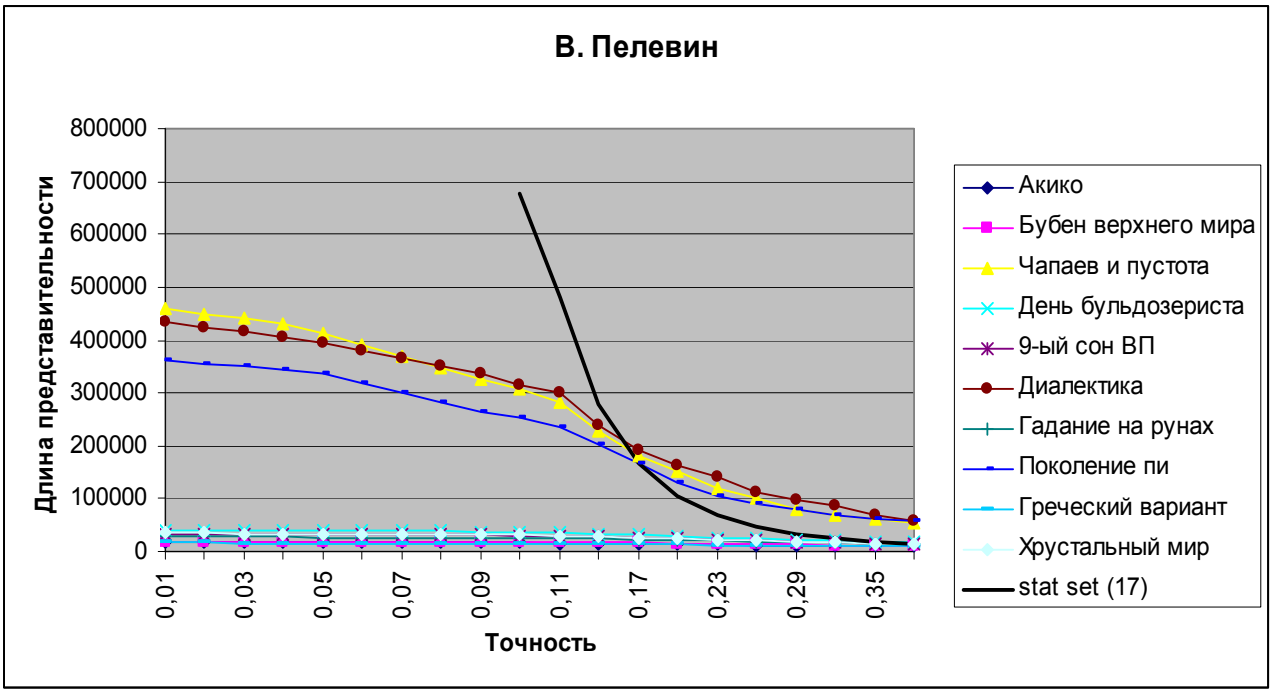

Рис. 5. Зависимость длины представительности от уровня стационарности для произведений В. Пелевина, 3-ПФР

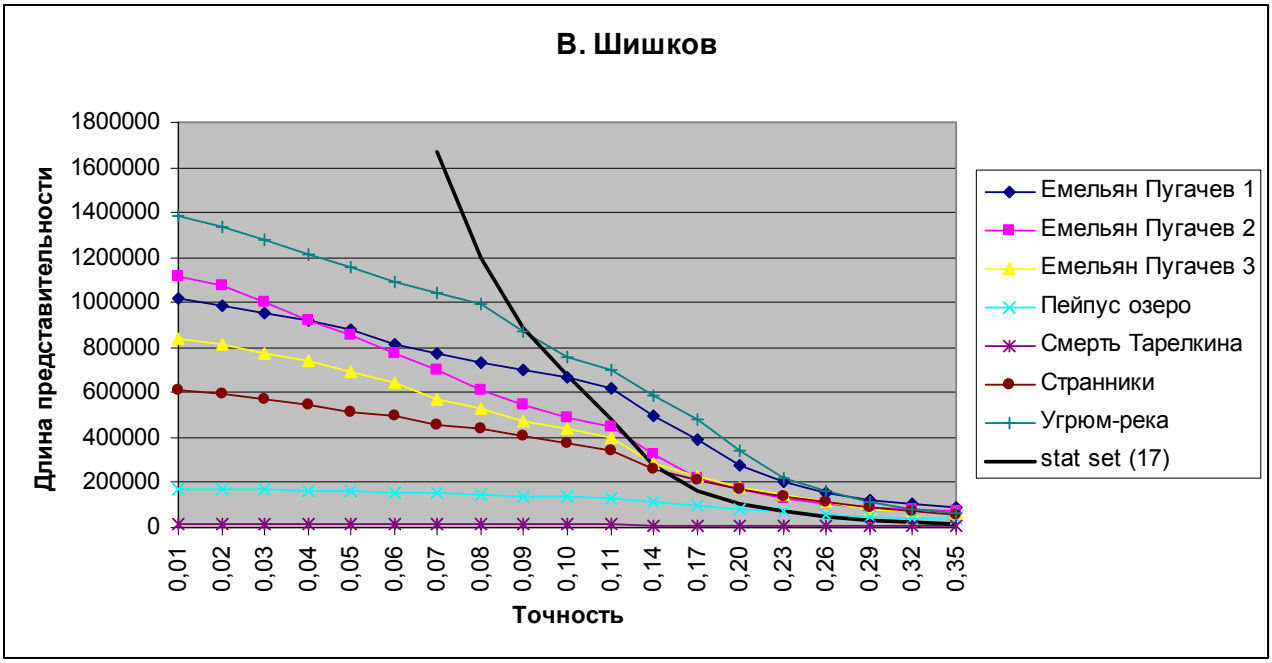

Рис. 6. Зависимость длины представительности от уровня стационарности для произведений В. Шишкова, 3-ПФР

Из графиков рис. 4-6 следует, что отрывок может отличаться от полного текста на величину, существенно превосходящую точность, оцененную по формуле (17). Кроме того, на малых длинах распределение не является стационарным с принятой точностью – той, с которой оцениваются вероятности. С другой стороны, как будет показано ниже, начиная с некоторой длины теста, приближение к ПФР полного текста происходит быстрее, чем это следует из формулы (19). Это означает, что при написании литературного произведения ПФР формируется под действием как авторских предпочтений в выборе последовательности слов, так и объективных факторов, таких, как жанр или формат произведения. В то же время из рис. 4-6 следует, что длина представительности не является индикатором писателя, как и жанра.

Конечно, чем больше произведение, тем больше и длина представительности для одного и того же уровня стабилизации. Тем удивительнее факт распознавания авторов текстов малых длин (10-15 тыс.) методом, который по статистическим критериям вроде бы не должен давать точных результатов. Видимо, это связано с эффектом стабилизации ПФР на длинах, существенно меньших, чем длина полного текста.

Как видно из приведенных примеров, длина представительности больших текстов при низкой точности лежит выше, чем кривая стационарной длины «stat set», отвечающая этой точности и определяемая по формуле (17). Тексты малой длины быстро выходят на асимптотический режим в соответствии с формулой (19), и потому могут считаться квазистационарными. Если же длина текста превышает 400 тыс. знаков, то 3-ПФР относительно малых фрагментов текста имеет нестационарное поведение. Оно отвечает области, расположенной

справа от точки пересечения кривой «stat set» с кривой длины представительности.

Например, различающиеся почти на порядок по длинам тексты В. Шишкова все укладываются в 0,25-стационарность на длинах менее 200 тыс. знаков, однако, если бы текст в целом играл бы роль генеральной совокупности для своего фрагмента, должна была бы наблюдаться как минимум 0,17-стационарность. Заметим также, что произведения В. Пелевина удовлетворяют последнему требованию, т.е. 0,17-стационарны на длине 200 тыс., а тексты Ю. Никитина 0,11-стационарны на длине 400 тыс. Тем не менее, на меньших длинах фрагментов текстов этих авторов уровень стационарности (17) не достигается.

С другой стороны, слева от точки пересечения кривой «stat set» с кривой длины представительности тексты уже стационарны. Поскольку же длина фрагмента в этой точке существенно короче, чем это могло бы быть по формуле (19), то ПФР текста стабилизируется в силу авторской специфики письма, а не из-за конечной длины произведения. На рис. 7 приведены длины представительности для совокупности 60 текстов близких длин разных авторов.

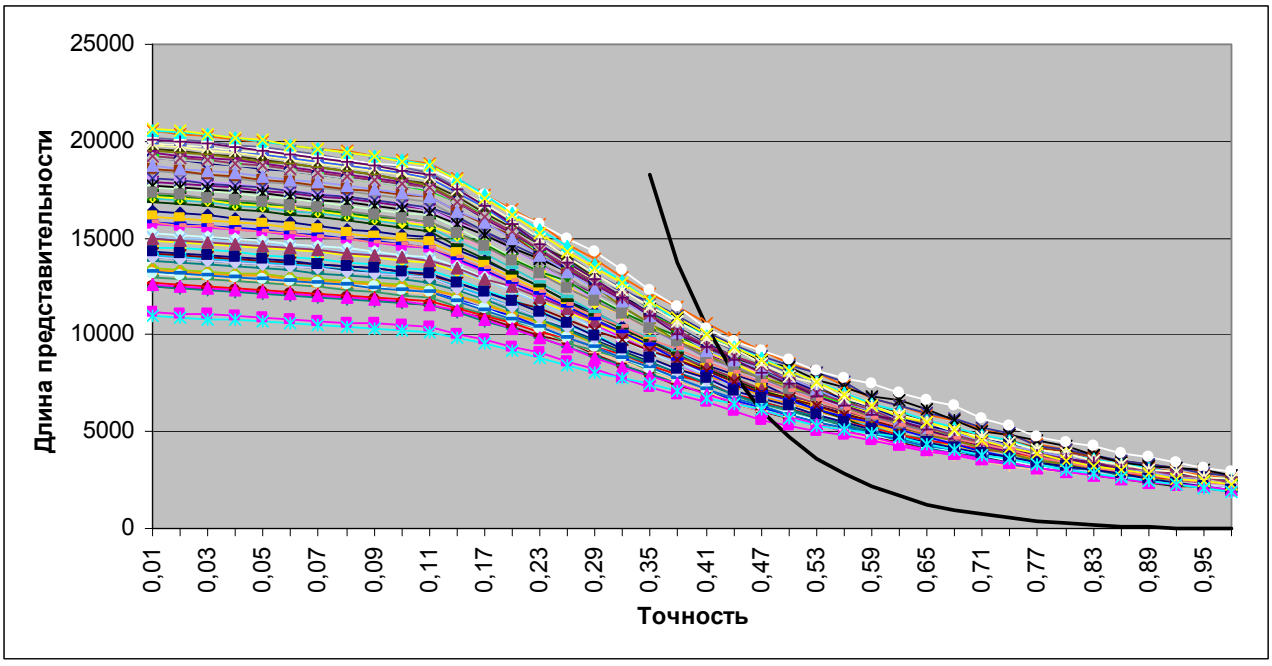

Рис. 7. Длина 3-ПФР представительности текстов 11-21 тыс. знаков

На графиках рис. 7 хорошо просматривается конечный линейный участок длин представительности, отвечающий формуле (19). Видно также, что кривая минимальной длины текста (17) пересекает эти длины где-то в средней части. Анализ 300 текстов 30 разных писателей, длины которых (текстов) лежали в диапазоне 7 тыс. – 1 млн знаков, с высокой точностью подтверждает это частное наблюдение. Зависимость точки пересечения кривой «stat set» с кривыми длин представительности показана на рис. 8.

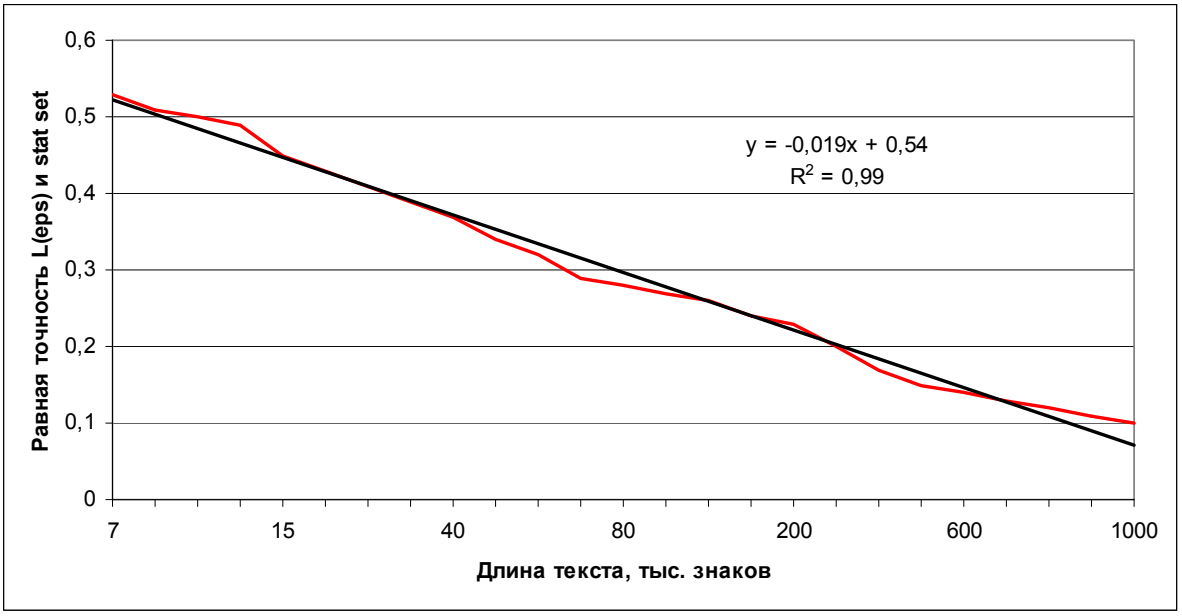

Рис. 8. Абсцисса (точность) точки пересечения кривой «stat set» и $L(\varepsilon)$

Высокая детерминация линейной аппроксимации графика на рис. 8 свидетельствует о том, что абсцисса точки пересечения линейно зависит от длины текста. Линейность означает, что стационарность 3-ПФР достигается на одной и той же части произведения независимо от его длины. Эта доля в процентах численно равна абсолютной величине котангенса угла наклона аппроксимирующей прямой на рис. 8, т.е. приблизительно $1/0{,}019 \approx 52{,}6\%$. Отсюда, в частности, следует, что две половины одного и того же текста отличаются одна от другой по своим 3-ПФР примерно на 2ε, где ε есть уровень стационарности, достигаемый на длине представительности, равной длине стационарной оценки (17). Это вывод подтверждается на практике.

Таким образом, кажущаяся нестационарность 3-ПФР начальных фрагментов текста объясняется тем, что буквы появляются в тексте не полностью случайно, а представляют собой последовательность зависимых величин. Видимо, последнее обстоятельство и приводит к тому, что идентификация автора по его эталонной 3-ПФР является наиболее точным методом из прочих, проанализированных в [1].

Распределения меньших размерностей (1-ПФР и 2-ПФР) имеют другое поведение. Как показано в [1], 1-ПФР большинства текстов стабилизируются на уровне, достаточном для идентификации автора в соответствии с данными табл. 1 и 2. Этот уровень не зависит от длины полного текста и составляет примерно 30 тыс. знаков для всех писателей, чему отвечает точность оценки вероятностей появления букв в тексте, равная 0,05. Иными словами, $L(0{,}05) \approx 30000$. Аналогично и 2-ПФР большинства текстов стабилизируется на уровне точности, примерно равном 0,11, так что $L(0{,}11) \approx 80000$. В то же время точки пересечения длин представительности с кривой точности (17) плавающие, не имеющего такого четкого поведения, как для 3-ПФР.

5. Статистический эксперимент определения автора текста

Для иллюстрации эффективности метода приведем фрагмент таблицы результатов эксперимента по определению автора текста в соответствии с методом (4-6) близости 3-ПФР по выборке «30 авторов, 300 текстов». Абсолютно во всех случаях каждый из десяти текстов уверенно сопоставлялся своему автору по принципу наименьшего расстояния до соответствующего авторского эталона. Для примера мы показываем определение авторов лишь некоторых классических текстов (Гоголь, Достоевский, Тургенев).

Были выбраны следующие произведения Н.В. Гоголя: 1 – «Вечера на хуторе…, ч.1», 2 – «Вечера на хуторе…, ч.2», 3 – «Вий», 4 – «Иван Иванович и Иван Никифорович», 5 – «Мертвые души, ч.1», 6 – «Мертвые души, ч.2», 7 – «Невский проспект», 8 – «Старосветские помещики», 9 – «Портрет», 10 – «Тарас Бульба». Произведения И.С. Тургенева: 1 – «Вешние воды», 2 – «Дворянское гнездо», 3 – «Дым», 4 – «Записки охотника» (целиком), 5 – «Накануне», 6 – «Новь», 7 – «Отцы и дети», 8 – «Первая любовь», 9 – «Рудин», 10 – «Степной король Лир». Также были выбраны произведения Ф.М. Достоевского: 1 – «Бесы», 2 – «Братья Карамазовы» т.1, 3 – «Братья Карамазовы» т.2, 4 – «Братья Карамазовы» т.3, 5 – «Братья Карамазовы» т.4, 6 – «Записки из мертвого дома», 7 – «Идиот», 8 – «Подросток», 9 – «Преступление и наказание», 10 – «Униженные и оскорбленные».

Результаты идентификации приведены в табл. 3 (а, б, в), где показаны расстояний от испытуемых текстов «1-10» до авторских эталонов.

Табл. 3 (а). Пример идентификации автора по 3-ПФР (Гоголь)

Автор\текст	1	2	3	4	5	6	7	8	9	10
Айтматов	0,45	0,43	0,49	0,57	0,41	0,42	0,52	0,47	0,49	0,44
Акунин	0,48	0,46	0,52	0,54	0,37	0,40	0,47	0,43	0,47	0,48
Булгаков	0,48	0,47	0,52	0,55	0,40	0,44	0,50	0,45	0,49	0,48
Гоголь	**0,35**	**0,31**	**0,39**	**0,44**	**0,19**	**0,27**	**0,41**	**0,29**	**0,37**	**0,32**
Донцова	0,52	0,50	0,57	0,60	0,43	0,48	0,55	0,50	0,53	0,54
Достоевский	0,48	0,44	0,51	0,50	0,35	0,37	0,46	0,38	0,44	0,49
Маркеев	0,52	0,52	0,56	0,62	0,45	0,48	0,55	0,51	0,52	0,51
Набоков	0,49	0,48	0,51	0,56	0,38	0,42	0,45	0,43	0,44	0,48
Толстой Л.Н	0,49	0,46	0,50	0,54	0,36	0,38	0,46	0,40	0,44	0,46
Тургенев	0,44	0,42	0,50	0,52	0,35	0,38	0,47	0,41	0,46	0,46

Табл. 3 (б). Пример идентификации автора по 3-ПФР (Достоевский)

Автор\текст	1	2	3	4	5	6	7	8	9	10
Айтматов	0,41	0,43	0,43	0,43	0,43	0,40	0,44	0,43	0,41	0,45
Акунин	0,35	0,40	0,42	0,40	0,40	0,41	0,40	0,39	0,37	0,45
Булгаков	0,39	0,44	0,47	0,43	0,44	0,44	0,44	0,44	0,41	0,48
Гоголь	0,35	0,37	0,39	0,38	0,39	0,35	0,38	0,38	0,35	0,41
Донцова	0,42	0,47	0,49	0,45	0,47	0,47	0,46	0,45	0,42	0,47
Достоевский	**0,18**	**0,23**	**0,24**	**0,25**	**0,22**	**0,30**	**0,17**	**0,17**	**0,18**	**0,25**
Маркеев	0,45	0,49	0,51	0,48	0,50	0,47	0,50	0,49	0,46	0,54
Набоков	0,39	0,43	0,46	0,44	0,43	0,41	0,43	0,42	0,41	0,47
Толстой Л.Н	0,36	0,39	0,41	0,41	0,40	0,38	0,37	0,37	0,37	0,38
Тургенев	0,31	0,36	0,39	0,36	0,38	0,38	0,34	0,34	0,33	0,36

Табл. 3 (в). Пример идентификации автора по 3-ПФР (Тургенев)

Автор\текст	1	2	3	4	5	6	7	8	9	10
Айтматов	0,42	0,44	0,41	0,37	0,42	0,41	0,42	0,45	0,47	0,45
Акунин	0,37	0,40	0,36	0,34	0,39	0,37	0,38	0,45	0,45	0,42
Булгаков	0,38	0,42	0,40	0,35	0,41	0,39	0,41	0,46	0,48	0,44
Гоголь	0,38	0,41	0,37	0,31	0,38	0,36	0,38	0,44	0,44	0,42
Донцова	0,42	0,44	0,43	0,39	0,43	0,42	0,43	0,48	0,49	0,48
Достоевский	0,37	0,39	0,33	0,36	0,35	0,33	0,34	0,42	0,37	0,42
Маркеев	0,45	0,47	0,45	0,40	0,48	0,45	0,47	0,51	0,53	0,49
Набоков	0,37	0,41	0,37	0,36	0,41	0,38	0,40	0,44	0,46	0,44
Толстой Л.Н	0,38	0,39	0,36	0,38	0,39	0,37	0,37	0,43	0,41	0,45
Тургенев	**0,26**	**0,26**	**0,24**	**0,22**	**0,25**	**0,22**	**0,23**	**0,33**	**0,30**	**0,33**

Аналогично выглядят фрагменты таблицы и для произведений других авторов: «свой» текст всегда оказывается ближе к соответствующему авторскому эталону. Следует заметить, что если ввести ограничение на идентификацию автора по уровню разделения $\hat{\rho}$, то получится ошибка, равная 0,03: десять текстов из 300 не найдут своего автора. Например, при идентификации текстов Н.В. Гоголя обнаружилось, что тексты «3», «4», «7» и «9» довольно далеко отстоят от авторского эталона (больше расстояния разделения из табл. 1), но от эталонов других авторов они отстоят еще дальше. Следовательно, ограничение $\rho \le \hat{\rho}$ может порождать ошибочные заключения, как, впрочем, и любое другое статистическое условие, избранное на роль экспертной оценки. Отметим, что методом близости 3-ПФР без ограничения $\rho \le \hat{\rho}$ удалось безошибочно идентифицировать тексты такого многоликого писателя, как Л.Н. Толстой.

Приведем несколько статистических аргументов в пользу эффективности метода 3-ПФР по сравнению с распределениями другой размерности. Во-первых, ошибка второго рода, являющаяся в задаче идентификации ключевой, минимальна для 3-ПФР (см. табл. 1). Во-вторых, для распределений расстояний от текстов до «своих-чужих» эталонов (рис. 1) площадь перекрытия графиков также минимальна для 3-ПФР. В-третьих, если вычислить разность средних расстояний от текста до «чужого» и до «своего» авторов $l_d - l_s$ (см. табл. 1 и рис. 1) и сравнить ее с суммой соответствующих дисперсий $\sigma_d + \sigma_s$, то для 1,2,4-ПФР отношение $\dfrac{l_d - l_s}{\sigma_d + \sigma_s}$ равно 0,9, а для 3-ПФР оно равно 1,1, т.е. различающая способность авторских эталонов здесь наибольшая.

Важно подчеркнуть, что увеличивать размерность ПФР для повышения точности анализа имеет смысл лишь до определенных пределов. Обе последовательности характерных расстояний между текстами – «своими» и «чужими» – возрастают с увеличением размерности ПФР, но они также и ограничены сверху одним и тем же числом, ибо эти расстояния заведомо не превосходят двойки. Следовательно, точность, с которой могут быть различены тексты разных авторов, при увеличении размерности ПФР начнет уменьшаться, что потребует для анализа очень больших объемов текстов. Это наблюдается уже на 4-ПФР.

Идентификация некоторых текстов может быть сделана более надежной, если у автора достаточно произведений, написанных в разных жанрах, и тогда можно было бы рассмотреть вместо одного несколько авторских эталонов. Для этого надо предварительно кластеризовать тексты автора. Следуя [1], мы будем кластеризовать произведения методом попарной близости ПФР на уровне, не превосходящем определенного расстояния ρ^*. Именно, в один кластер попадают те и только те произведения, попарные расстояния между которыми не превосходят ρ^*. Если некоторый текст оказался одновременно в двух и

более кластерах, то в целях использования «четкой логики» он считается принадлежащим тому кластеру, в котором больше элементов. Если число элементов таких кластеров одинаково, то выбирается тот из них, для которого дисперсия расстояния от данного текста до остальных элементов кластера минимальна.

Строго говоря, оптимальное расстояние ρ^*, на котором следует проводить разделение кластеров, зависит от состава авторов. Смысл оптимальности состоит в минимизации ошибки разделения текстов разных авторов. На выборке «30 авторов – 300 текстов» минимальная ошибка, равная 0,15, получается при $\rho^* = 0{,}23$ для 2-ПФР, что превосходит расстояние разделения $\hat{\rho}$ из табл. 1. Использование распределений других размерностей приводит к большей ошибке. Так, для 1-ПФР при оптимальном значении $\rho^* = 0{,}065$ ошибка составила 0,25. Такая же ошибка 0,25 получается и для 3-ПФР при $\rho^* = 0{,}40$. Заметим, что оба оптимальных уровня кластеризации в двух последних случаях также больше расстояния разделения авторских эталонов. Это естественно, поскольку дисперсия распределения расстояний между «своими» текстами не меньше, чем дисперсия расстояний между текстами и «своим» эталоном (см. рис. 3).

Таким образом, для целей классификации текстов по жанрам (или авторов по «лицам») использовать 2-ПФР предпочтительнее, чем 3-ПФР. Следовательно, задачи идентификации и кластеризации наилучшим образом решаются разными инструментами: 3-ПФР и 2-ПФР соответственно. Это связано с тем, что 3-ПФР учитывает более тонкие различия между авторами, чем 2-ПФР, что повышает вероятность распознавания «своих» текстов, но препятствует точной кластеризации. Причем методы также различны: для идентификации лучше анализировать расстояние от текста до эталона, а для кластеризации – расстояния между отдельными текстами.

Такие «моно-писатели», как Тургенев, Набоков, Достоевский, имеют один кластер, куда входят почти все их произведения, за исключением одного-двух. Их «нетипичные» тексты не относятся и к другим писателям, а стоят в библиотеке особняком. Вопрос о том, следует ли их учитывать при составлении эталона автора, или лучше исключить, не имеет однозначного ответа. Наш эксперимент показал, что на уровне 3-ПФР кластеризация авторов по «лицам» не имеет значения для идентификации их произведений.

Многогранные авторы большую часть своих произведений пишут в различных стилях. Например, 10 текстов Гоголя (см. табл. 3-а) объединяются на уровне 0,40 для 3-ПФР в два кластера: «1, 2», «5, 6, 8», а остальные тексты стоят особняком. Та же кластеризация по 2-ПФР на уровне $\rho^* = 0{,}23$ дает более компактную картину: имеется один кластер «1, 2, 5-10», и два отдельно стоящих текста – «Вий (3)» и «Иван Иванович … (4)».

Все вышеприведенные примеры относятся к группе так называемых писателей профессиональных, и для них предложенный метод показал высокоточный результат. Однако в практических приложениях может возникнуть и несколько иная задача – идентификация автора, не являющегося в прямом смысле профессиональным писателем. В этой связи полезным представляется рассмотрение в данной работе «нетипичного писателя», совокупный объем произведений которого достаточен для возможности применения статистического метода. В произведениях такого автора от произведения к произведению редко сохраняется какая-то определенная стилистика, а также часто присутствует большее разнообразие жанров. Все это несколько выводит задачу за пределы первоначально очерченной области, однако аргумент здесь тот, что в своей области метод показал высокоточный результат, и можно надеяться, что он даст ответы на интересующие вопросы и в другом случае, который окажется достаточно близким по постановке задачи.

Таким «непрофессиональным писателем» в нашем примере является Елена Ивановна Рерих (1879-1955), творческое наследие которой весьма разнообразно: дневники и путевые заметки, научные статьи и художественные произведения, письма и книги религиозного содержания. В следующем разделе по описанной методике будет проведен анализ ее работ с демонстрацией статистического метода в решении известной проблемы авторства цикла «Живая Этика».

6. Анализ литературного наследия Е.И. Рерих

Литературное наследие Е.И. Рерих (далее Е.И.) по данным Международного Центра Рерихов [6] можно разделить на четыре группы:

- письма;
- переводы (с английского): книги Е.П. Блаватской «Тайная доктрина» (два первых тома из трех), а также избранные письма Махатм, объединенные в сборник «Чаша Востока»;
- собственные произведения: «Основы буддизма», «Криптограммы Востока», «Знамя преподобного Сергия Радонежского», сборник статей «Огонь неопаляющий» и короткий очерк «Три ключа»;
- цикл из 14-ти книг «Живая Этика» (иначе Агни Йога).

В последнюю группу включены книги, изданные в 1924-1938 годах. Их интересная особенность состоит в том, что по просьбе самой Е.И. все они вышли без указания автора. Е.И. неоднократно объясняла это тем, что истинным автором этих работ являлся некий Махатма (иначе «Великий Учитель») Мориа, существование которого, впрочем, многие подвергали сомнению уже тогда [7]. По словам самой Е.И., она лишь записывала его мысли, общаясь с ним путем «яснослышания» и «автоматического письма». На Западе, где фамилия Рерихов широко известна, уже с момента начала публикации первых книг цикла началась дискуссия по поводу их возможного

авторства. Критически настроенные исследователи видели во всем этом мистификацию и замаскированное мошенничество, определенно считая Е.И. автором этих книг. Не вдаваясь здесь в обсуждение возможности яснослышания как способа передачи информации, равно как и саму принципиальную возможность существования подобных Махатм, сформулируем основной вопрос нашего статистического исследования: можно ли объединить две последние группы в один однородный кластер? Если – да, то, скорее всего, все эти произведения написаны собственно ею без посторонней помощи. Если – нет, то возникает задача кластеризации ее текстов с одновременным вопросом о том, кто мог бы быть ее фактическим соавтором.

Для начала кластеризуем первые две группы произведений – письма и переводы. Письма и перевод двух томов «Тайной доктрины» Блаватской четко отделяются статистическим методом в два кластера. В кластер с переводом этих двух томов попадает и перевод третьего тома Блаватской, выполненный, однако, уже другим переводчиком (А.П. Хейдок). Это совпадает с тем наблюдением [1], что переводчик не является соавтором переводимого текста, и автор в переводах своих текстов на другие языки четко определяется как самостоятельный писатель независимо от переводчика (переводчиков). В этой связи интересно определить место переведенной Е.И. книги «Чаша Востока», поскольку, по уверениям авторов, издавших эту книгу на английском языке (это А.П. Синнет и А.О. Хьюм, английские чиновники в Индии в конце XIX в.), написана она была также «с помощью» Махатмы Мориа. Близка ли «Чаша» циклу «Живая Этика» или подгруппе каких-либо произведений из этого цикла?

Таким образом, объектом анализа являются следующие произведения, которые мы занумеруем для краткости последующих ссылок. Цикл «Живая Этика»: 1 – «Листы Сада Мории. Зов», 2 – «Листы Сада Мории. Озарение», 3 – «Община», 4 – «Знаки Агни Йоги» , 5 – «Беспредельность, ч.1», 6 – «Беспредельность, ч.2», 7 – «Иерархия», 8 – «Сердце», 9 – «Мир Огненный, ч.1», 10 – «Мир Огненный, ч.2», 11 – «Мир Огненный, ч.3», 12 – «Аум», 13 – «Братство, ч.1», 14 – «Братство, ч.2». Другие произведения: 15 – «Чаша Востока», 16 – «Основы буддизма», 17 – «Криптограммы Востока», 18 – «Знамя преподобного Сергия Радонежского». Сборник статей «Огонь неопалаяющий» содержит много цитат из писем Е.И. и принадлежит этому кластеру, а очерк «Три ключа» имеет малый объем и далее не рассматривается.

Оказалось, что все эти 18 текстов, как приписанные одному автору, отделяются от других литераторов по методу 3-ПФР (см. п.5). Это, конечно, не особенно впечатляет, поскольку тематика произведений Е.И. слишком узкая, и для корректной проверки близости ее текстов к авторскому эталону следует взять для сравнения что-либо похожее. Как уже говорилось, переводы Блаватской отделяются и не смешиваются с «Живой Этикой». В качестве второго автора, с которым интересно было бы сравнить тексты Е.И., мы взяли работы Н.К. Рериха (далее Н.К.): 1 – «Адамант», 2 – «Держава света», 3 – «Обитель света», 4 – «О вечном», 5 – «Врата в будущее», 6 – «Листы дневника,

ч.1», 7 – «Листы дневника, ч.2», 8 – «Листы дневника, ч.3», 9 – «Нерушимое», 10 – «Сердце Азии», 11 – «Химиват». Для этого имеется несколько соображений: во-первых, Н.К. был мужем Е.И. и мог активно участвовать в ее творческой и научной судьбе; во-вторых, налицо также схожесть литературных и религиозных интересов супругов.

Если сравнить эти два авторских кластера – Е.И. и Н.К., то получится, что все тексты Н.К. ближе к своему эталону, но два текста Е.И. – а именно, 17 и 18 – ближе не к Е.И., а к Н.К. Однако следует отметить, что эти ближайшие расстояния все же весьма велики для того, чтобы признать этого автора своим – 0,45 и 0,53 соответственно для текстов 17 и 18. При этом все произведения Н.К. удалены от авторского эталона на расстояние, не превышающее 0,36, а среднее расстояние от его текстов до эталона равно 0,22. Поэтому при ограничении $\rho \le 0{,}40$ следовало бы признать, что автор текстов 17 и 18 – другой человек, не Е.И. и не Н.К.

Однако с такими выводами надо быть аккуратным и не забывать о Гоголе (см. табл. 3-а), который тоже писал довольно разнообразно. Ведь именно эти тексты достоверно написаны самой Е.И. Следовательно, произведения Е.И. 16, 17, 18 имеет смысл выделить в отдельную группу, отделив их тем самым от «Живой Этики». Расстояния же между самими этими тремя текстами порядка 0,55-0,60, т.е. все они весьма различны и не объединяются в кластер по признаку попарной близости.

Распределение расстояний между 3-ПФР рассмотренных текстов Е.И. и Н.К. приведено на рис. 9. Из него следует, что Н.К. писал единообразно, тогда как Е.И. имеет распределение, не характерное для одного автора (ср. с рис. 3).

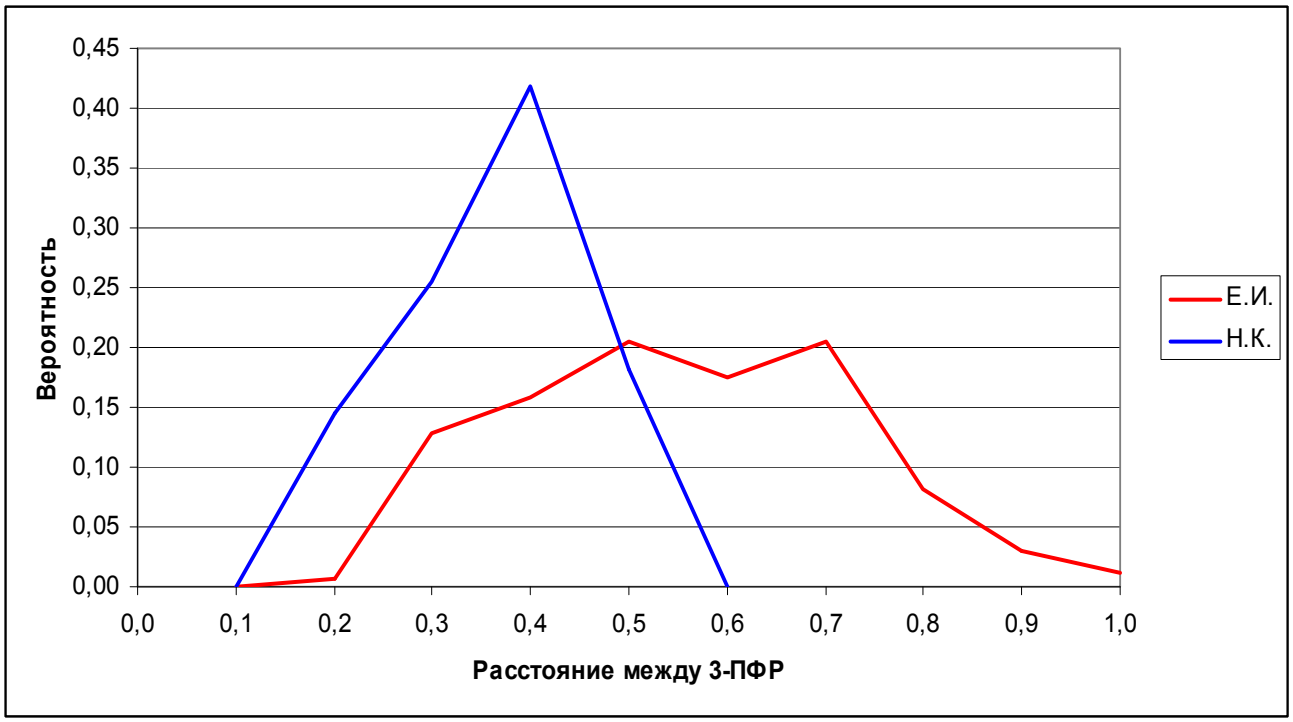

Рис. 9. Распределение расстояний между «своими» текстами Е.И. и Н.К.

Переходя теперь к циклу «Живая Этика» (тексты 1-14), следует отметить, что ситуация здесь тоже неоднородна, как и при рассмотрении произведений Е.И. в целом. Тексты 1, 5 и 6 имеют большие расстояния до эталона: 0,53, 0,48 и 0,63 соответственно. Остальные тексты написаны достаточно компактно со средним расстоянием до эталона, равным 0,25. Таким образом, эти три текста – «Листы сада Мории. Зов» и обе отдельно изданных части «Беспредельности» – написаны в другом стиле. Расстояние между текстами 5 и 6 равно 0,36, так что «Беспредельности» могут быть объединены в один кластер, но текст 1 стоит особняком: он далек абсолютно от всех произведений Е.И., а также и от Н.К. Ближайшим к нему текстом является текст 2, но и тот отстоит от него на расстояние 0,44. При этом «Чаша Востока» не принадлежит основному кластеру из текстов «Живой Этики», отстоя от эталона на 0,43.

Следует, впрочем, заметить, что ближе всего «Чаша» находится к эталону Блаватской (расстояние 0,31), которая также утверждала, что ее духовным наставником был Махатма Мориа. Более того, это произведение попарно объединяется со всеми тремя томами «Тайной доктрины» на уровне $\rho \le 0{,}41$. Это вполне укладывается в концепцию независимости первоначального автора от переводчика. Дело в том, что автор «Писем Махатм» А.П. Синнет был членом Теософского общества, одним из основателей и идейным вдохновителем которого была Блаватская, причем, по словам Синнета, она-то и «помогла ему вступить в духовную связь с Махатмами» [8]. Тем самым представляется наиболее вероятным, что «Письма…» написаны самой Блаватской.

Итак, кластеризация текстов Е.И. привела к образованию трех «лиц»: Е.И.1 включает в себя тексты «Живой Этики» 2, 3, 4, 7-14; Е.И.2 состоит из двух текстов 5 и 6; Е.И.3 содержит произведения 16-18; тексты 1 и 15 стоят особняком, но ближе всего из эталонов Е.И. оказываются все же к «Живой Этике». При таком разделении три текста 16-18, не будучи близки между собой, оказываются ближе всего к своему средневзвешенному эталону, чем к Н.К. Их расстояния до своего эталона равны соответственно 0,33, 0,45 и 0,29.

Из проведенного анализа можно сделать следующие выводы (с коррекцией на их вероятностный характер).

1. Н.К. не писал тексты, которые можно было бы трактовать как не вполне собственные произведения Е.И.

2. Если и существовал Махатма Мориа, то диктовал он свои мысли весьма изобретательно, ибо «Чаша» и «Листы сада Мории» не близки, а книги Агни Йоги не однородны по статистической структуре. Иными словами, Махатмы, с которыми мысленно общались Блаватская и Рерих – разные. «Письма Махатм», переведенные Е.И. с английского, скорее всего, были записаны А.П. Синнетом со слов Е.П. Блаватской.

3. Расстояния между текстами Е.И. очень большие, некоторые достигают даже величин 0,97 («Беспредельность, ч.2» и «Криптограммы Востока»). Это не то что нетипично для текстов одного автора, но даже и для разных авторов

встречается довольно редко. Следовательно, при написании книг Агни Йоги весьма вероятно совместное творчество. В этом смысле, скорее всего, справедливы утверждения Е.И. о своем взаимодействии с другими носителями культуры Востока при написании текстов цикла «Живая Этика».

По поводу последнего вывода следует еще раз напомнить, что, согласно самой Е.И., все книги цикла «Живая Этика» были созданы на основе ее общения с Махатмой Мориа. Поэтому следовало бы ожидать, что они будут находиться в одном кластере. Но в результате проведенного статистического анализа мы получили разложение этого цикла в три кластера, ни один из которых не близок Блаватской. Следовательно, можно допустить, что у Е.И. были по крайней мере два соавтора при написании этих текстов, и ни один из них не был «тем самым» Махатмой Мориа.

7. Заключение

В данной работе мы в основном сосредоточились на оценке точности распределений триграмм, чтобы объяснить их высокую идентифицирующую способность при анализе авторства. Действительно, их статистические свойства таковы, что позволяют наиболее точно распознать автора по сравнению с ПФР других размерностей. Вместе с тем выяснилось, что задача кластеризации лучше решается с помощью 2-ПФР методом попарной близости.

Описанные подходы могут быть применены и к задачам идентификации текстов по другим параметрам, которыми можно характеризовать произведения, кроме авторской принадлежности: тематическое направление, литературный жанр, эпоха написания и т.п. Для задачи идентификации автора первый подход (сравнение с эталоном) оказался более эффективен, чем второй. Это связано с тем, что ширина разброса расстояний между текстами одного автора несколько больше, чем ширина разброса расстояний между ними же и средней авторской ПФР. Для жанров метод среднего «жанрового эталона» может быть не особенно эффективен, потому что при большом количестве авторов, принимающих участие в формировании такого эталона, ПФР будет приближаться к некоторой типичной ПФР для всего языка как такового, и точность такого инструмента окажется низкой.

При анализе творчества Е.И. Рерих мы не имели целью решить некую культурологическую проблему. Задача стояла в демонстрации возможностей метода в решении вопроса, по которому у ряда исследователей нет единого мнения. Естественно, статистические выводы применительно к конкретной ситуации не могут служить доказательством верности той или иной точки зрения. Они лишь показывают, какой результат более вероятен, исходя из неких предпосылок, выполнимость которых, заметим, отнюдь не обязательна. На наш взгляд, главными особенностями всего, что так или иначе связано с Рерихами или Е.П. Блаватской, являются крайняя запутанность и историческая

неоднозначность. Так что в связи с этим обстоятельством и в виду возможных будущих фактологических уточнений эти предпосылки могут быть изменены.

Представляется, что перспективы применения статистических методов в литературе, языкознании и смежных областях весьма широки. Для их успешного использования нужен синтез усилий специалистов разных направлений.

Литература

1. Орлов Ю.Н., Осминин К.П. Методы статистического анализа литературных текстов. – М.: Эдиториал УРСС/Книжный дом «ЛИБРОКОМ», 2012. – 326 с.
2. Королюк В.С., Портенко Н.И., Скороход А.В., Турбин А.Ф. Справочник по теории вероятностей и математической статистике. – М.: Наука, 1985. – 640 с.
3. Орлов Ю.Н. Оптимальное разбиение гистограммы для оценивания выборочной плотности функции распределения нестационарного временного ряда // Препринты ИПМ им. М.В. Келдыша РАН. 2013. № 14. 26 с.
URL: http://library.keldysh.ru/preprint.asp?id=2013-14
4. Абрамовиц М., Стиган И.М. Справочник по специальным функциям. – М.: Наука, 1979.
5. Орлов Ю.Н., Осминин К.П. Определение жанра и автора литературного произведения статистическими методами // Прикладная информатика, 2010. Т. 26. № 2. С. 95-108.
6. Электронная библиотека Международного Центра Рерихов.
URL: http://lib.icr.su/
7. Jenkins P. Mystics and Messiahs. Oxford University Press, 2000, NYC. p.41-42.
8. Sinnet A.P. The occult world. – London, 1881.